数字化校对技术在数字出版中的应用

郑晓慧
（河南人民出版社，河南 郑州 450016）

摘 要：针对传统出版模式下文字校对存在校对出错误个数与真实错误个数相差较大，并且无法给出准确性更高的合理改正建议等问题。通过构建文本编辑错误校对模型、错误检测前的数据平滑处理、基于数字化校对技术的错误检测、文本编辑错误改正与出版等，开展数字化校对技术在数字出版中的应用研究。通过对比数字化校对与传统校对应用效果的方式证明，基于数字化校对技术的数字出版校对出错误个数与真实错误个数相同，并且能够给出提高准确率的改正建议。

关键词：数字化出版；数字校对；文本编辑；数据处理；应用效果　　**中图分类号：**G23　　**文献标识码：**A
文章编号：1671-0134（2022）05-126-03　　DOI：10.19483/j.cnki.11-4653/n.2022.05.039
本文著录格式：郑晓慧 . 数字化校对技术在数字出版中的应用 [J]. 中国传媒科技，2022（05）：126-128.

导语

当前，越来越多的网络信息技术出现，并应用于各个领域当中，同时在社会需求不断推动下，数字化的出版形式产生。数字出版是将网络信息技术作为技术支撑，通过更具网络化的传播渠道，实现传播、阅读和生产方式的数字化。数字出版在发展过程中，为了不断适应和完善，逐渐演变出了多种类型的出版方式。[1] 数字出版与传统出版相比更具交互性，并且传播速度更快，可拓展面更广，在极大程度上提高了人们对图书的阅读需求，也进一步充实了现有图书资源。但随着数字出版发展速度的不断提升，在为其带来创新的同时，也使得诸多问题产生，例如数字出版信息数据量成倍增加，对校对、编辑等都造成巨大的负担。[2] 为了进一步探究数字化校对技术在数字出版当中的应用及应用效果，本文开展下述研究。

1. 数字化校对技术在数字出版中的应用

1.1 构建文本编辑错误校对模型

为了提高数字出版的质量，解决文本内容在编辑中出现错误的次数，本节提出一种针对文本编辑错误的校对模型。假设在编辑文本内容时，语句中文本内容表示为 S，则 S=S1，S2，S3…Sn，其中 1~n 表示构成文本内容的多个字节，在此基础上，采用全局检索的方式，对其中容易存在混淆的文字进行矩阵构建。并使用数字编辑设备中的统计功能项，进行全局参数的宏观调控，确保对编辑空间内文本数量统计结果的真实性与有效性。[3] 为了确保文本编辑错误校对模型在使用中的有效性，可在圈定检索空间后，使用文字统计法，进行混淆集合的人工识别与校对，人工操作编辑界面后，输出错误项集合，并使用文本中的替换功能，进行修正内容的重新校正，以此种方式，确保文本内容中所有校正的内容与局部修正需求匹配。但在此过程中应注意的是，在改正错误时，使用标注进行混淆文本的标记，并重点关注此部分文本内容的错误是否完全进行了修订，以此实现对文本编辑错误的有效校对。

根据上述论述，在明确文本编辑错误校对模型的基本需要后，设计如图 1 所示的模型总体框架。

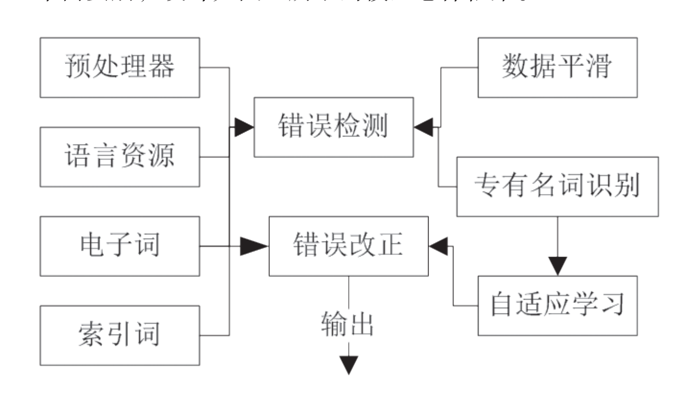


图 1　文本编辑错误校对模型总体框架结构图

从图 1 中文本编辑错误校对模型总体框架结构可以看出，检测错误部分输入的目标为需要进行校对的文本字符信息串，输出的结果为可能存在文本错误的位置。[4] 当将需要进行校对的文本字符信息串输入到构建的文本编辑错误校对模型当中时，根据局部文本的上下文语境，将可能存在错误的文本进行划分，并将该区域作为后续错误检测的重点位置区域。在对真实存在错误的文本进行改正后，再返回到上一阶段完成对错误检测结果的报告生成，并给出相应的改正建议。

1.2 错误检测前的数据平滑处理

按照本文上述论述内容，完成对文本编辑错误校对模型的构建后，为了确保后续错误检测的精度，在检测前还需要对数据进行平滑处理。由于需要进行校对的文本当中存在多种不同的错误成分类型，并且存在错误词语的位置上，其左右相邻的文本会出现数据稀疏的问题，上述问题的存在会造成检测难度增加，因此从多个方面实现对数

据的平滑处理。[5] 首先，针对文本窗口缩小的问题进行数据平滑处理。图 2 为文本窗口数据稀疏现象示意图。

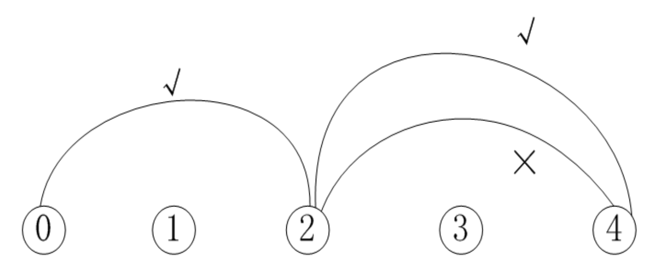


图 2　文本窗口数据稀疏现象示意图

图 2 中“×”符号表示为在文本窗口当中前后三个文字对出现了稀疏问题，“√”符合表示为文本窗口当中前后三个文字对未出现稀疏问题。从图 1 中所示的内容可以看出，若需要进行校对的文字当中其字符信息串 0~1 是按照正确的方式出现，而字符信息串 0~2 在文本窗口当中出现了数据稀疏问题，则说明 2 本身是一个存在错误的词语或 2 本身是正确的词语，但与 0~1 字符信息串连接后存在错误。[6] 针对上述存在问题，对其进行数据平滑处理，其计算公式为：

$$S_n = a \times y_n + (1-a) \times S_{n-1} \quad (1)$$

公式（1）中，a 表示为平滑系数；y_n 表示为在某一时刻 n 下，文本窗口平滑处理后的数据值；S_{n-1} 表示为在前一时刻通过平滑处理后的数据值；S_n 表示为经过平滑处理后的数据值。根据上述公式，针对图 1 当中存在的稀疏问题进行平滑处理，在处理的过程中，将第一次输入的原始文本数据作为初始状态数值，或将前几次输入的原始数据值的平均值作为初始状态数值。

其次，再对聚类词进行数据平滑处理。根据以往数字出版语言使用的经验得出，在文本当中存在很多同义词或近义词，通过其相互之间的转换，句子本身几乎不会存在差异，例如“观”和“看”、“认识”和“知道”等。[7] 通过近义词之间的相互转换，可以达到对文本数据平滑处理的效果。在进行平滑处理的过程中，还可引入同类词预料的方法，例如如下公式（2）表示同类词集：

$$N(1,2,\cdots,n) \prec N(1,2,\cdots,X_{ji},\cdots,n) \quad (2)$$

公式（2）中，N 表示为需要进行校对的目标文本；X_{ji} 表示为文本当中某一字符 i 的同类词集。通过上述操作，对文本窗口缩小和聚类词进行数据评价处理后，能够确保后续错误检测的准确度不受影响，提高数字化校对技术的应用性能。

1.3　基于数字化校对技术的错误检测

在检测前还需要将彼此容易混淆的词语进行收集，并形成混淆集合。在一个混淆集合当中包含了容易在使用过程中与校对目标词出现混淆的词语。在错误检测的过程中，引入一个分配器，用于对文本当中不同词语进行分类。在分类器进行过程中能够，对适合上下文语义的词语将其取值设置为 1，针对不适合上下文语义的词语，将其取值设置为 0。每个分配器都与文本上下文特征相关联，并且为每一个关联对象设置不同的连接权值。针对需要进行校对的目标词语进行获取，并在该词语上下连接的文本当中提取特征，将所有特征进行汇总，并得到如公式（3）所示的表达结果：

$$\theta = 1 \Leftrightarrow \sum_{f \in F} wf > \varepsilon \quad (3)$$

公式（3）中，θ 表示为利用分配器进行分类后得到的结果；F 表示为提取到的特征集合；w 表示为分配器判定结果数值，W 的取值为 0 或 1；f 表示为特征集合中的某一特征数值；ε 表示为分类常数。在错误检测过程中，所有连接的权值均为分配器通过多次学习获得的。因此，权值的学习可以看作是分配器判定错误的时候对取值进行调整的动态过程。根据学习过程中，不同类型分类器的实际表现，为其赋予不同的可行度权值，并将其带入到上述构建的文本编辑错误校对模型当中，实现对错误文本的检测。

1.4　文本编辑错误改正与出版

首先，从最小编辑距离角度出发，无论是在对自然语言进行理解还是处理的过程中，都会出现两个字符之间的距离问题，这种距离与普通意义上的距离不同，是指语义距离或编辑距离。在进行文本编辑错误改正过程中，通过对两个字符之间的最小编辑距离进行调整，可以实现对其改正。假设某一字符信息串为 A，其长度对应为 a，另一字符信息串为 B，其长度对应为 b，则此时 A 和 B 之间的编辑距离为 ed（A[a]，B[b]）。在进行改正的过程中，编辑操作会引起“时间”问题产生，需要一定的“时间”才能够缩短两个字符信息串之间的编辑距离。在改正中，通常设定一次的编辑改正操作需要使用单位 1 的“时间”，一次才能够将编辑距离的“时间”量的计算等价转换为字符信息串编辑操作的次数，方便对错误改正次数的记录。

还可以通过易混淆集构建的方式，对文本编辑错误进行改正。将所有具有与被校对词语在某一特征上存在相似的不同词语汇总，并构成一个易混淆集合。这种特征可以是词语本身含义的相同，也可以是形或音等某个方面上的相同。通过对文本编辑错误进行观察，通常情况下产生的文本错误是由于文本当中正确词语被其相应的易混淆集合当中的词语所代替。因此，为了将其修改为正确的词语，将易混淆集合作为重要的候选词语集合。由于文字数量较大，因此易混淆集合在构建时难度较高，为了降低构建难度，利用现有词典附录扩充的方式构建易混淆集合，以此在易混淆集合的基础上完成对文本编辑错误的改正。按照上述内容将完成改正后的文本输出，并通过人工校对的方式，对其进行二次校对和三次校对，最终将完成校对的文本汇总，构成最终出版时的图书类型，以此完成对图书的校对和出版。

2. 数字化校对与传统校对应用效果分析

为了探究数字化校对技术应用后的数字出版与传统出版方式相比是否具备更高的应用优势，本文选择以某个图书的原始稿件作为研究对象，分别通过两种出版方式下的校对方法，对原始稿件进行校对，并记录两种校对方法的应用效果。在实验过程中，将原始稿件当中的所有文字内容设置为开放完全测试集，该集合当中包含了 200 个错误用例，记录两种方法校对得到的真实错误数量以及合理给出改正建议的个数，并通过计算得出改正建议的准确率。由于两种校对方法在实际应用中计错误个数方式不同，为了确保实验结果的公正性，对其错误文字计数标准进行规定：首先，针对同一页面当中反复出现的错误文字，最多标记为四个错误个数；其次，针对扉页上出现的文字错误，最多标记为两个错误个数；最后，针对文章当中存在影响语义、不符合版面要求的文字或需要空格而未空格的错误，每处计 1 个错误个数。按照上述错误文字计数标准，记录两种校对方法的校对结果，并绘制成如表 1 所示的结果。

表 1　数字化校对与传统校对应用效果对比

真实错误个数 / 个	数字化校对		传统校对	
	校对出错误个数 / 个	给出合理改正建议个数 / 个	校对出错误个数 / 个	给出合理改正建议个数 / 个
50	48	48	32	29
100	99	99	69	66
150	149	149	102	98
200	200	200	162	131

从表 1 中记录的实验数据可以看出，尽管真实错误个数为 50 个时的校对出错误个数为 48 个，但随着校对真实错误个数的增加，数字化校对能够对之前完成的校对内容进行反复检查，因此能够确保将最终所有 200 个真实错误个数全部检测出来。但传统校对方法在完成对之前内容的校对后，不会对其进行反复检查，因此最终造成校对出错误个数与真实错误个数相差较大的问题产生。数字化校对能够实现对所有开放完全测试集中错误内容的标记，并给出相应的改正意见，而传统校对方法校对出错误个数相比较少，并且无法针对已经发现的校对错误给出相应的改正意见。通过进一步对两种校对方法的改正建议准确率计算得出，数字化校对的准确率高达 100%，而传统校对方法的准确率仅为：131 ÷ 200 × 100%=65.5%。因此，通过上述实验及得出的实验结果可以证明，数字化校对方法在应用到数字出版当中时，能够实现对所有错误内容的准确校对，并给出准确率更高的改正建议。将该技术应用到数字出版当中，可进一步促进出版行业向着数字化、信息化的方向发展。

结语

数字化校对技术不仅可以应用在出版领域中，还可应用于各类文字处理领域当中，未来随着数字化校对技术的不断完善，其校对应用性能也将逐渐提升，从最基础的自动分词，到语义语法分析等。尽管当前数字化校对技术的应用仍然处于刚刚起步的阶段，未来还会遇到更大的困难和挑战。从当前研究水平来看，仍然存在几方面问题需要解决。例如，当前数字化校对受到错误实例缺少等多种条件限制；基于长词模糊匹配对校对技术进行优化等。在今后研究中，还将针对上述存在问题进行更加深入研究，从而进一步提高数字化校对技术的应用性能。媒

参考文献

[1] 胡雪婵 . 数字出版知识服务中的传统文化表达及应用——兼谈汉语语料库中的汉语成语语义韵特点 [J]. 出版广角，2021（16）：80−82.

[2] 韦林枝 . 国家数字出版基地管理运行模式探析——以开发区主导的“园中园”发展模式为例 [J]. 新闻潮，2021（8）：41−44.

[3] 孟良荣，许鹤平 . 高校图书馆电子书服务模式思考——基于亚马逊数字出版平台构建的启示 [J]. 内蒙古科技与经济，2021（17）：75−77.

[4] 原业伟 . 战“疫”和脱贫两项出版工作的年终总结——第十届中国数字出版博览会线上数字内容精品展侧记 [J]. 新阅读，2021（2）：7−8.

[5] 张新新 . 中国特色数字出版话语体系初探：实践与框架——2020 年中国数字出版盘点 [J]. 科技与出版，2021（3）：86−97.

[6] 何蓉 . 合同法视域下数字出版著作权问题研究——以法国出版合同改革为借鉴 [J]. 科技与出版，2021（5）：110−114.

[7] 翁晓峰 . 失衡与治理：学术期刊数字出版产业链利益分配问题研究 [J]. 中国科技期刊研究，2021（8）：1016−1025.

作者简介：郑晓慧（1977−），女，河南新密，中级编辑，研究方向：图书校对。

（责任编辑：胡杨）